Modern Alchemy: Neurocognitive Reverse Engineering

Olivia Guest[1,2], Natalia Scharfenberg[1,2], and Iris van Rooij[1,2,3]

[1]Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
[2]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands
[3]Department of Linguistics, Cognitive Science, and Semiotics & Interacting Minds Centre, Aarhus University, Denmark

The cognitive sciences, especially at the intersections with computer science, artificial intelligence, and neuroscience, propose 'reverse engineering' the mind or brain as a viable methodology. We show three important issues with this stance: *1*) Reverse engineering proper is not a single method and follows a different path when uncovering an engineered substance versus a computer. *2*) These two forms of reverse engineering are incompatible. We cannot safely reason from attempts to reverse engineer a substance to attempts to reverse engineer a computational system, and vice versa. Such flawed reasoning rears its head, for instance, when neurocognitive scientists reason about what artificial neural networks and brains have in common using correlations or structural similarity. *3*) While neither type of reverse engineering can make sense of non-engineered entities, both are applied in incompatible and mix-and-matched ways in cognitive scientists' thinking about computational models of cognition. This results in treating mind as a substance; a methodological manoeuvre that is, in fact, incompatible with computationalism. We formalise how neurocognitive scientists reason (*metatheoretical calculus*) and show how this leads to serious errors. Finally, we discuss what this means for those who ascribe to computationalism, and those who do not.

Keywords: mutliple realisability, reverse engineering, computational cognitive neuroscience, metatheoretical calculus, computationalism

> We can only presume to build machines *like us* once we see ourselves as *machines first.*
>
> Abeba Birhane (2022, p. 13)

Cognitive science — especially subareas that bring into contact psychology's and neuroscience's (hyper)empiricist methods with those of computer science, such as cognitive computational neuroscience — introduces us to multitudinous potential reasoning pitfalls (e.g. Guest & Martin, 2023, 2024, 2025; van Rooij et al., 2024). In the best case scenario, these issues are carefully and cautiously dealt with, tempered, and sometimes avoided. In the worst cases, these reasoning traps are set off, ensnaring scholars, who in their work attempt to characterise brain, behaviour, and cognition. Once trapped, these scholars end up reasoning in ways that violate their own theoretical (computational or otherwise) commitments and principles. An overlooked pitfall we dedicate this paper to is the implicit or explicit, as well as inconsistent, use of 'reverse engineering' in the cognitive sciences. Altogether, the lack of attention to this pitfall results in the current status quo wherein "the literature contain[s] 'curious shadowy' (Russell, 1918) syllogisms that will never obtain" (Guest & Martin, 2023, p. 219). Specifically, we will show that metatheoretic syllogisms — our reasoning over apparent reverse engineering attempts — exist in an environment that will never be truth-making.

The claim that cognition is or can be scientifically understood as computation, i.e. *computationalism*, is a serious stance and not a throw-away statement. A commitmemnt to computationalism forces certain ways of reasoning about cognition and rules out others (Chirimuuta, 2021; Egan, 2017; Guest & Martin, 2021, 2023; Hardcastle, 1995, 1996; van Rooij, 2008; van Rooij et al., 2024). Words such as 'computational' (to refer to e.g. models or theories) and 'computationalism' (to refer to e.g. scientists' stance or philosophical commitments), and even everyday nouns like 'computer' and verbs like 'compute', are technical terms that bind us to stringent meanings and formal concepts (also see Guest & Martin, 2025, Box 1). If one believes that the brain and/or mind are a type of computer, then one cannot in good faith use reverse engineering of substances (as opposed to of computers) as a methodology. Yet, as we demonstrate herein, this occurs often and consistently within modern cognitive scientific reasoning.

Olivia Guest https://orcid.org/0000-0002-1891-0972

Author note: The kernel of the idea for this paper appeared when Olivia was semimindlessly watching QI Series 20 XL: Secrets, Spies & Sleuths. Sandi Toksvig presents the panellists with a ballpoint pen and explains that China, despite being the world's biggest producer of them, was importing the tips until 2017 because they did not know how to make them, especially with respect to the steel used in the ball. On hearing this, Daliso Chaponda says "I don't see how you can reverse engineer an iPhone, but not a ballpoint pen!" The first author thought to herself: What an excellent question; I do see how.

The authors would like to thank members of the Computational Cognitive Science lab at Donders Centre for Cognition for useful discussions and feedback, and especially Nils Donselaar for feedback and suggestions for some aspects of the formalisms used in this paper. Any remaining errors are our own.

Correspondence: Olivia Guest, Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands. E-mail: olivia.guest@donders.ru.nl

In this paper, we provide the reader with formalised lenses that can be applied to scientists' reasoning to determine if commitment to reverse engineering as a methodology is at play (Table 1), what kind of thesis about the brain or mind is implied by the type of their reverse engineering (for substances or for computers; Table 2), and what can be done to disentangle such commitments. The case is made herein as to why reasoning over reverse engineering is a real challenge for the psychological, neuro-, and cognitive sciences — heretofore unresolved, but not unresolvable. To presage our conclusion, even if we grant reverse engineering is an appropriate methodology (cf. Guest & Martin, 2023; Marom, 2009; Rueckl, 2012; Schierwagen, 2012; van Rooij et al., 2024), scientists may not shuttle between treating cognition as a substance and as a computer willy-nilly because utterly different inferences follow directly from the two commitments (cf. Hardcastle, 1995; Leibniz, 1714; Rozemond, 2014; Ryle, 2009/1949; Sayward, 1983).

1.1 Overview

> Paralogisms and antinomies are the evidence that an expression is systematically misleading.
>
> Gilbert Ryle (1931, p. 168)

We construct a *metatheoretical calculus* — a formal exposition of the reasoning from theory to model to system under study and back (Guest, 2024; Guest & Martin, 2023, 2024) — to unpick how, as modern computationalists, we have become unstuck from our own stated beliefs and have fallen into unwanted and unnecessary paradoxical disarray. On the one hand, we reason as if mind is substance; and on the other as if it is a computer. In this paper, we explain how to detect and avoid what we see as serious reasoning and methodological issues going forwards (viz. Stebbing, 2022). Specifically, we define and analyse:

a) what reverse engineering is both in the original true case and as adopted by cognitive scientists, even when not explicitly labelled (in section 2: **Cognitive science as reverse engineering**);

b) how true reverse engineering manifests in two divergent ways using case studies and formal treatments: in section 3, **Reverse engineering substances**, where we discuss Chinese porcelain as reverse engineered by European alchemists and Jesuit missionaries; and in section 4, **Reverse engineering computers**, where we discuss IBM's System 360 as reverse engineered by the USSR and the Eastern Bloc;

c) how in our (neuro)cognitive scientific reasoning reverse engineering of substances in its appeal to correlation and structural similarity appears as if it provides the right frame (in section 5: **Mind-as-substance versus mind-as-computer**); relatedly,

d) how cognitive (neuro)scientists treat mind as substance or as computer, mixing and matching to suit rhetorical ends in violation of multiple realisability, of functionalism, and of computationalism (also in section 5: **Mind-as-substance versus mind-as-computer**); and finally,

e) how to address this, by being open about our beliefs about mind and brain, and aware of what they bind us to (in section 6: **Will the real computationalist please stand up?**)

2 Cognitive science as reverse engineering

> Our adventure is actually a great heresy. We are about to conceive of the knower as a computing machine.
>
> Warren S McCulloch (1954, p. 143)

Reverse engineering is seen as a cognitive scientific methodology by a non-negligible number of practitioners and philosophers (e.g. Cauwenberghs, 2013; Chater & Brown, 2008; Chater et al., 2010, 2011; Dawson, 2013; Denić & Szymanik, 2022; Dennett, 1995; DiCarlo, 2018; Dupoux, 2018; Griffiths et al., 2024; Gurney, 2009; Harnad, 2003; Harnad, 2025; Haspel et al., 2023; Hurley et al., 2013; Jackson et al., 2021; Jonas & Kording, 2017; Lake et al., 2016; Levshina, 2021; Marom, 2009; Miłkowski, 2013; Paul et al., 2023; Pietraszewski & Wertz, 2011; Schierwagen, 2012; Schrimpf et al., 2021; Tenenbaum, 2021; Tenenbaum et al., 2011; Yoo et al., 2024; Zednik & Jäkel, 2016). "For decades, *reverse engineering* the brain has been one of the top priorities of science and technology research." (Yoo et al., 2024, p. 2) It is, both explicitly and implicitly, embraced, especially by modern connectionists (who depend on deep artificial neural networks for their scientific modelling; e.g. DiCarlo, 2018; Dupoux, 2018; Haspel et al., 2023; Schrimpf et al., 2021; Zednik, 2018; for discussion and critique see: Guest and Martin, 2023, 2024; van Rooij et al., 2024).

In the general and original case, outside the cognitive sciences, reverse engineering is "the process of extracting know-how or knowledge from a human-made artifact" according to Pamela Samuelson and Suzanne Scotchmer (2001, p. 1577; others propose similar definitions, e.g. Aplin, 2013; Eilam, 2011). In the cognitive sciences and specifically in the cognitive computational neuroscientific setting, in the best case scenario, reverse engineering is described as analysing the neurocognitive system to derive specifications and then given those specifications proposing functional and/or mechanistic solutions that fulfil these criteria (cf. Chirimuuta, 2013, 2021; Darden, 2007, 2008; Egan, 2017; Guest & Martin, 2021; Millikan, 1989; Stinson, 2018; Sullivan, 2022; van Rooij & Wareham, 2008). But some go further. As quoted by Marcin Miłkowski (2013) to motivate his treatment of reverse engineering in cognitive science, Dennett (1995) claims that "[r]everse engineering is [...] the interpretation of an already existing artifact by an analysis of the design considerations that must have governed its creation." (p. 683) Dennett goes on to make some extreme claims:

> What Marr, Newell, and I (along with just about everyone in AI) have long assumed is that this method of reverse engineering was the right way to do cognitive science. Whether you consider AI to be forward engineering (just build me a robot, however you want) or reverse engineering (prove, through building, that you have figured out how the human mechanism works), the same principles apply.
>
> And within limits, the results have been not just satisfactory; they have been virtually definitive of cognitive science. That is, what makes a neuroscientist a cognitive neuroscientist, for instance, is the acceptance, to some degree, of this project of reverse engineering.
>
> (1995, p. 684)

He asserts that reverse engineering is inexorably interwoven into the fabric of our field.[1] Many others also follow suit, e.g. "Cognitive science is, after all, a process of reverse engineering" (Chater et al., 2010, p. 812). Some go further still, proclaiming tellingly that adopting this methodology is essentially not scientific, suggesting that cognitive scientists:

> act like engineers trying to reverse-engineer the human mind. [...] Instead of defining the research and theorizing side of psychology as a science, we can define it as a form of reverse engineering.
>
> Thomas Leahey (2005, pp. 139–140)

Sometimes this methodology is deployed implicitly, i.e. without the literal phrase 'reverse engineering', but nonetheless we can uncover it is used by how the work is carried out. Frequently, however, it is outlined explicitly:

> One of the most widespread research strategies in computational cognitive neuroscience is a top-down (or "reverse- engineering") strategy[, i.e. "to begin by answering questions at the computational level and to work downwards".] These questions are answered by specifying a mathematical function that describes the system's behavior[.]
>
> [R]everse-engineering is a matter of inferring the function and structure of mechanisms from (among others) prior characterizations of the behavioral and cognitive phenomena for which they are deemed responsible.
>
> Carlos Zednik (2018, pp. 358–359)

In the connectionist flavour of reverse engineering, deep artificial neural network models are deployed that (learn to) map inputs (e.g. pixels of photorealistic images) to outputs (e.g. vectors that represent to the human user linguistic labels of said images; Guest and Martin, 2024; Shiffrin and Mitchell, 2023). And proponents of this approach ask questions such as: "Could these models also let us reverse-engineer the brain mechanisms of higher-level human cognition?" (Schrimpf et al., 2021, p. 1). The Bayesian conception of cognition also condones reverse engineering in tandem with its connectionist fellow traveller (e.g. Chater & Brown, 2008; Chater et al., 2010, 2011; Griffiths et al., 2024; Lake et al., 2016; Paul et al., 2023; Schierwagen, 2012; Tenenbaum et al., 2011).[2] Strikingly, the functional analysis (the specification; Guest & Martin, 2021) appears completely absent. It also appears that this methodology and related stances and questions are in tension.

To examine what all this entails, let us grant that reverse engineering is a possible useful methodology. And let us answer: What are the properties of a metatheoretical position that espouses reverse engineering as a cognitive scientific methodology? Van Rooij et al. (2024) class this stance under *makeism*, which takes computationalism to imply: *a*) cognition can be (re)made in computational systems; *b*) (re)making cognition implies we can explain and/or understand it; and/or *c*) explaining and/or understanding cognition has (re)making it as a prerequisite. Many examples of makeist views can be found in mainstream cognitive science wherein "the (re-)construction of the original system is attempted by creating duplicates including computer models" (Schierwagen, 2012, p. 145). And in the quote above from Dennett (1995): "prove, through building, that you have figured out how the human mechanism works" (p. 684). Additionally, even stronger beliefs are present: "[t]he only way to make sure that we understand a mechanism and have its complete causal model is to replicate the mechanism in a different medium." (Miłkowski, 2013, p. 19) This family of stances results from "viewing the mind as a highly complex computational device, [and thus] it becomes natural to think of cognitive science as a process of 'reverse engineering' — or more specifically, 'reverse computer science"' (Chater & Brown, 2008, p. 37). For the non-makeist computationalist, the cognitive system can be understood as computing, or explained computationally, without needing to recreate it (e.g. Chirimuuta, 2021; Polger & Shapiro, 2023; van Rooij et al., 2024). For the makeist, on the other hand, the mind (or brain) is seen as an artifact that computes, and one that can be (reverse) engineered.

In the following sections, we furnish the reader with two types of true reverse engineering: in section 3, **Reverse engineering substances**, and in section 4, **Reverse engineering computers**. Based on the forthcoming analyses, it is possible to tease out, formally contain, and compare and contrast relevant scientific reasoning patterns, called metatheoretical calculi (Guest, 2024; Guest & Martin, 2023). A metatheoretical calculus forces us to confront what our commitments entail, elucidating in this case where computationalism points one way, and in fact rules out any other way, but practitioners fail to notice.

3 Reverse engineering substances

3.1 Porcelain

Perhaps unexpectedly, porcelain — or rather the search for the recipe (method plus ingredients) to produce it — is a very useful example for understanding the cognitive endeavour of reverse engineering a substance in practice. Here we do not discuss the discovery of porcelain by Chinese experts, but the reverse engineering of porcelain by the West once they knew Chinese experts had discovered it.

"About 20,000 years ago, hunter-gatherers living in a cave 170 kilometres northeast of Jingdezhen [in Jiangxi province, China] made the oldest-known pottery in the world" (Gillette, 2016, p. 11). Therefore, it is perhaps not surprising that porcelain was invented in China "in a primitive form during the Tang dynasty (618–907) and in the form best known in the West dur-

[1] A common thread with many who propose that cognitive science and reverse engineering are required, or defined, to be deeply entangled accounts is that they appear to depend on evolutionary theorising, including evolutionary psychology in the case of cognition (e.g. Boudry & Pigliucci, 2013; Csete & Doyle, 2002; Dennett, 1995; Dretske, 2000; Erneling & Johnson, 2005; Figdor, 2023; Harnad, 2003; Jonas & Kording, 2017; Leahey, 2005; Simon, 1995). This may raise alarm bells for some (viz. Guest, 2024).

[2] Possibly, this need to deploy engineering strategies, like reverse engineering, is exacerbated because of the inherent intractability of their Bayesian models (viz. Kwisthout et al., 2011; van Rooij & Wareham, 2012). Evidence for this can be found in statements such as:

> when scaled-up to real-world problems, full Bayesian computations are intractable, an issue that is routinely faced in engineering applications. From this perspective, the fields of machine learning, artificial intelligence, statistics, informational theory and control theory can be viewed as rich sources of hypotheses concerning tractable, approximate algorithms that might underlie probabilistic cognition.
>
> Nick Chater et al. (2006, p. 290)

ing the Yuan dynasty (1279–1368)" (Encyclopædia Britannica, 2024, n.p.). Jingdezhen was the epicentre of porcelain production, where farmers first made these ceramic wares during their "agricultural slack season [using a raw material] called china stone [that] produces ceramics that are extremely white, pure, and translucent." (Gillette, 2016, p. 4).

According to Maris Boyd Gillette (2016), Europeans came into contact with porcelain from Jingdezhen in the fifteenth century.

> [However, s]mall numbers of vessels trickled into Europe during the thirteenth and fourteenth centuries[. And began to arrive in larger quantities once direct maritime trade routes between the Indian Ocean and the Mediterranean were established at the beginning of the sixteenth century. [...] By the eighteenth century, Chinese porcelain could be found in most households across Europe, while potters in the West still struggled to discover [its] secrets.
>
> Alejandra Gutiérrez et al. (2021, p. 1213)

This has even been pushed back further with examples found in Almería, Spain dated from the ninth and eleventh centuries (Gutiérrez et al., 2021).

> From the sixteenth through to the early eighteenth centuries, [Europeans] succumbed to 'porcelain disease', purchasing massive amounts of Jingdezhen ceramics (Finlay, 2010). Jingdezhen porcelain was such a valuable commodity that European governments sent spies to discover how it was made. Entrepreneurs across Europe tried to replicate porcelain using media such as glass, crystal, soapstone, barium, and animal bones.
>
> Maris Boyd Gillette (2016, p. 22)

Perhaps surprisingly, it took until about 1575 for the first European soft-paste porcelain to be made in Florence — and further still till 1708 for hard-paste or true porcelain to be produced in Saxony by Johann Böttger, an alchemist (Gillette, 2016). On the industrial espionage side, a Jesuit missionary, François Xavier d'Entrecolles (1712, 1722) wrote two letters back to France detailing the specific ways Jingdezhen porcelain was made (Leonard, 2006; Vashisth & Kumar, 2013). Finally, a third type of porcelain (bone china) was discovered in 1794 by Josiah Spode in the United Kingdom, whose company still produces porcelain to this day (Encyclopædia Britannica, 2024; Özgündoğdu, 2005; see also rows *Equivalence sought*, *Multiple realisation*, and *Solution instances* in Table 1). As the name suggests, it uses bone ash obtained from cattle bones, to increase this type of porcelain's "white and semi-transparent characteristics[, making it] whiter, thinner and more transparent than other porcelain types." (Özgündoğdu, 2005, p. 30)

Together these events, reverse engineering through alchemy and industrial espionage (row *Search strategy* in Table 1), meant that in the last three decades of the eighteenth century Europeans produced their own porcelain with no dependence on Chinese imports (Kerr and Wood, 2004). True porcelain then was first created by Europeans roughly four centuries after it was invented by Chinese experts, and at least one hundred years after it was widespread in Europe (row *Search duration* in Table 1).

This historical example, serves as a useful guide to thinking about parts of collective human cognition, that manifests as true reverse engineering.[3] In this case, the example involves reverse engineering of a substance. This engineering task can be summarised as: given a known instance of the substance (e.g. porcelain), find a way of creating a close enough replica (e.g. soft-paste, hard-paste or true, or bone china). As reviewed in Cognitive science as reverse engineering, for many this is analogous to how cognitive science ought to function. This characterisation of modern cognitive science as construing mind-as-substance may appear shocking to the reader. However, seeing scientists' reasoning this way will reframe and explain many aspects of the disagreements and paradoxes within and between modern computationalist thought. To facilitate transparant analysis of the conceptual implications of reverse engineering mind-as-substance, we next present a formalisation of this type of reverse engineering.

3.2 Formally searching for substances

> Reverse-Engineering Substances
>
> *Given:* A set of possible ingredients I, a class of possible methods $\mathcal{M}$, and a target substance T, defined as a subset of properties[4] $\{t_1, t_2, ..., t_l\} = T \subseteq P$.
>
> *Task:* Find a subset of ingredients $I' = \{i_1, i_2, \dots i_n\} \subseteq I$ and a method $m : 2^{|I|} \rightarrow 2^{|P|}$ of type $\mathcal{M}$, such that m applied to I' yields a substance $m(I') = S = \{s_1, s_2, ..., s_k\} \subseteq P$ that is *structurally equivalent* to T (denoted $S \approx T$). Here, two substances S and T are said to be structurally equivalent if and only if all (relevant) properties that T posseses also S possesses, i.e. for each $t \in T$ there exists an $s \in S$, such that $s \approx t$.

Reverse engineering a substance can formally be conceptualized as a search in a combinatorially complex space.[5] Above, we present a (semi-)formal characterisation of how one can possibly think about reverse engineering of a substance as a kind of search in a complex space of possibilities. This formalism serves merely to improve conceptual transparancy and to facilitate comparison and contrast for when we consider reverse engineering a computer (vs a substance). For readability, the formalism leaves many details unspecified, but suffices for our purposes.

[3] Analogous to how cognitive science is a form of collective human cognition and how cognitive science is conceived of as reverse engineering by makeism (recall section 2, **Cognitive science as reverse engineering**).

[4] Our formalisation of 'substance' is generally consistent with 'bundle theory' (i.e. the idea that substances are just bundles of properties; see Gyekye, 1973 for a critique). However, we make no deep metaphysical commitment about the nature of substances, and here merely assume these properties can be defined by humans insofar as relevant to their purposes.

[5] To illustrate just how combinatorially complex this search space is, consider that the number of possible subsets I' grows exponentially in the number of possible ingredients to choose from (i.e. $2^{|I|}$). Even for 50 possible ingredients, there already exist more than 150,000,000,000,000,000,000 possible mixtures. Moreover, the class M may allow for all kinds of methods of mixing ingredients in various quantities and performing actions on the mixtures (such as heating, cooling, pressuring, centrifuging, steaming, etc.). Since operations can be applied in different orderings and even repeatedly on subsets of the sets of ingredients, in sequence and in parallel, the number of possible methods is vast as well. Given the astronomical size of $2^{|I|}$ and the vast number of possible candidates for m neither can be assumed to be 'given' to the reverse engineer in any explicit sense, but can only be intensionally specified.

Table 1: Comparison of the differences between two historical cases of true reverse engineering. In the case of porcelain, for example, the candidate substance (an artifact that does not compute; row *1*) must be highly correlated on material properties, if not identical to, porcelain (rows *2*, *3*, *6*, & *7*). In contrast, for the case of reverse engineering a computer, internal and external parts may diverge completely, having no relationship with each other — there are infinitely many (potential) solutions (rows *3* & *6*). In both cases, to both limit the search space (for both it is unbounded) and search time (row *4*) complexities, peeking at the solution (row *5*) was a necessary part of reverse engineering.

Reverse Engineering Proper		
	Porcelain, substance	**ЕС ЭВМ, computer**
1) *Computation performed*	none or identity function	(universal) Turing machine
2) *Equivalence sought*	structural	functional
3) *Multiple realisation*	minimally or uniquely realisable	massively or infinitely realisable
4) *Search duration*	thousands or hundreds of years	single digit number of years
5) *Search strategy*	industrial espionage, alchemy	industrial espionage, engineering
6) *Solution instances*	very few, one to three	infinite
7) *Verification method*	correlation	principles of computation

Given some target substance S, the reverse engineering task is to find a subset of ingredients $I' = \{i_1, i_2, \ldots, i_n\} \subseteq I$ and a method m, such that when method m is run on input I' it produces a substance S (row *Computation performed*, Table 1). Conceptually, we exclude from I any singleton sets, since we are looking for how to make S, and we are not interested in finding S ready-made. This also excludes from $\mathcal{M}$ methods that map the input to the output directly, as this would be akin to having S ready-made.

We can check whether we have found a valid pair (m, I') when S is produced and compared to a set of target properties, $T = \{t_1, t_2, \ldots, t_l\}$, which it can be verified using correlation or identity (row *Verification method*, Table 1). So $m(I') = S$ can be verified as a solution by examining the properties of $S = \{s_1, s_2, \ldots, s_k\}$ by checking if $S \approx T$. For each substance and target property pair, $s \in S$ and $t \in T$, different methods can be used to check for equivalence, but all can be defined as easy and quick to evaluate to make this problem (marginally) simpler. In other words, correlation or equality can be relatively easy to compute, such that knowing if T and S are instances of the 'same' substance is readily knowable once S is produced by a given (m, I') pair (rows *Multiple realisation* and *Equivalence sought* Table 1).

Given these properties it might be the case that if a given S appears to match our T but not closely enough (if we set certain deviation bounds for properties), then we may be getting close to finding a neighbouring S' that does match T. Here, 'neigbouring' means that there is only a small change in the chosen set or quantity of ingredients or only a small change in the method. In other words, if, say, T requires the substance to be light-permeable to a certain degree and S is close but outside our transparency target, then it might stand to reason that S' might make the cut. While there is no guarantee for this, it might be a useful strategy when searching the space of possibilities.

Other useful strategies that can diminish the vastness of this search space — although again with no guarantee — can be to condition testing a given m if it includes firing the ingredients in a kiln after mixing them together. One might think, mistakenly, that it is therefore a requirement for finding a good m to fire up a kiln and then conditional on that run a variety of methods using a kiln. But industrial espionage can be part of the search for m, which involves no direct use of a kiln. In other words, a good m can be constructed through a successful attempt at industrial espionage which obtains the steps to a recipe for porcelain. Thus requiring no firing up of a kiln in the search for an m, only in the process of applying $m(I')$ to obtain S since we need the substance itself to check if it is porcelain. Other so-called heuristics might be used too, for example: the ingredients should be as white as the finished product, but one of the unfired ingredients "ought to border a little on the green side" (d'Entrecolles, 1712; also: Kerr and Wood, 2004); the ingredients should be mined from the ground, but recall animal bone ash is used to make bone china, which "has the highest translucency property in standard product thicknesses among all the porcelain types.' (Özgündoğdu, 2005, p. 30); or that the unfired wares when placed in the kiln should retain their shape and size, unlike baked goods which notoriously deform and expand, since the thickness and opacity of the finished product are important. Historically, a mix of both industrial espionage and laborious firing of pottery in a kiln was necessary for discovering porcelain (see *Search strategy*, Table 1).

The search for porcelain describes a type of reverse engineering, when it comes to (re)creating a known substance given a sample of the substance or given its material properties. In the case of porcelain, this gave rise to (at least) three types of porcelain (row *Solution instances*, Table 1), with the recipe (method plus ingredients) for true porcelain being obtained through both laborious trial and error as well as industrial espionage, i.e. stealing the recipe (row *Search strategy*, Table 1). This may be unsurprising given the vastness of the search space — even if shrunk using common sense, i.e. requiring a kiln to be part of the instructions, it is still huge (see footnote 6) — but we will see in the following sections how this vastness somehow gets left behind (recall this took hundreds of years; row *Search duration*, Table 1) when our reasoning contains (perhaps un-

wittingly) the assumption that mind is substance. Which is to say, we cannot steal the recipe for the brain or mind, so how are we even imagining we can meaningfully shrink this space?

Before we move to the next section, Reverse engineering computers, it must be noted that when a given S is compared to the target T, the process can be such that there are infinite properties for comparison; properties need not be limited to a fixed length, l. For a given substance, S, which we suspect is porcelain, when we compare it to true porcelain from Jingdezhen, T, we can keep generating data points. In other words, we may keep asking if $s_i \approx t_j$ while i and/or j go to infinity (or some very large value), by continuously sampling various measurements from the two materials. In fact, if indeed it is the case that S is structurally equivalent to T (within the bounds we define) this continuous sampling and checking should prove to generate more and more confirmatory evidence that they are the 'same' for our purposes. So, whether l is small, large, or infinite, the more properties that S matches on T, the more certain we can be that we have hit on two instances of the same substance, that we have found the right recipe, $m(I') \approx T$. As we move down our list ticking off properties, and they match, we can, with each match, get more excited that we have "cracked the porcelain code" (Leonard, 2006). And vice versa, we know *a priori* that true porcelain will match on all these properties with itself. Clearly, these properties hold for the actual case of deciding if two substances match (rows *Equivalence sought* and *Verification method*, Table 1). Therefore, one may think that it stands to reason that a substance will match with itself. To this we say: yes, substances have this property by definition. Computers, Turing Machines, do not.

4 Reverse engineering computers

4.1 EC ЭВМ

In this section, while we remain within the scheme of reverse engineering, we move to an example and paradigm that involves computation, i.e. the target of our search is a digital computer (row *Computation performed*, Table 1; Fyrbiak et al., 2017; Rekoff, 1985). This puts us in line with how computationalists who espouse reverse engineering and therefore makeism should reason — in line with how their metatheoretical calculus should tick along — in principle. While no formal treatment of these two cases in the way we describe has been given, others outside cognitive science have noticed a distinction:

> A knife can easily be reproduced by knowing its dimensions, its materials, and how those materials were treated. A microprocessor probably cannot be reproduced from a specimen or collection of specimens; in all likelihood it is easier to start with a clean sheet of paper. In fact, there are a large number of people who are not convinced that it is possible to perform a "complete" test on a microprocessor chip at all.
>
> Michael G. Rekoff (1985, pp. 244–245)

The EC ЭВМ (also known as the ES EVM, Unified System, or Ryad) series of computers were clones, i.e. faithful duplicates, of the IBM System 360 series of mainframes (Central Intelligence Agency, 1973; Gray & Smith, 2001; Levin, 2016a; Szabó, 2020; Пржиялковский, 1997; Савватеев, 2023). The System 360 was announced in 1964 and appeared on the market in 1965, promising upwards and downwards compatibility between the series' machines' firmware and software, i.e. they created an abstraction layer that allowed the same machine code to run on all models within the series (Amdahl et al., 1964; Watson & Petre, 2000). Cloning these computers was intended to help the USSR and the Eastern Bloc generally take advantage of inter alia the breadth of software that ran on IBM computers (Davis and Goodman, 1978; Donig, 2010; Levin, 2016a, 2016b; Szabó, 2020; row *Equivalence sought*, Table 1). What the Soviet and other Council for Mutual Economic Assistance (Comecon) reverse engineers did to achieve this, similar to the Western attempts to recreate porcelain, was industrial espionage as well as making use of legally obtained licences, mixed with — not alchemy this time![6] — their knowledge of the principles behind computer hardware and software engineering (Goodman, 1979; Levin, 2016a; Malinovsky et al., 2010; Szabó, 2020; row *Search strategy*, Table 1).

Going back to the start of the story in the 1960s, "companies such as IBM [...] promulgated their concepts through teaching materials[, thus] ultimately reshaping these patterns of exchange. Eastern Europe, like Western Europe, at that time lacked a center of comparable gravity" (Donig, 2010, p. 34). With the advent of IBM's System 360 mainframes in 1965, which were incredibly well-received by the Western world (Gray & Smith, 2001; Watson & Petre, 2000), many companies in, e.g. Germany, Britain, the USA, and so on, made their own clones. For example, the RCA's (a company based in the USA) Spectra 70 brochure boasted that:

> all non-privileged instructions, formats and character codes are identical with the corresponding features in IBM's System 360. Yet, though they appear to be similar to the outside world when it comes to performing a specified job, internally the Spectra 70 systems are quite different, employing a uniquely individual logic, and exploiting a faster responsiveness which their special characteristics make possible.
>
> RCA Corporation (1965, p. 9)

Notably, they lean in to multiple realisability (Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2023; Hardcastle, 1995, 1996; Litch, 1997; Miłkowski, 2016; Polger & Shapiro, 2016; Ross, 2020). This underlines that the clone's internal components give rise to differences such as, e.g. faster hardware when compared to the original IBM System 360 (rows *Equivalence sought*, *Multiple realisation*, and *Solution instances*, Table 1).

The USSR, like the companies mentioned above who made clones, noticed IBM's successful System 360, with a 1966 report documenting that "a number of foreign publications emphasiz[e] its 'revolutionary character'." (Levin, 2016a) And so in 1967, inspired by IBM and undeterred by the embargoes imposed on it by the USA (Malinovsky et al., 2010; Yasuhara,

[6] The successes of alchemy and chymistry (direct precursor to modern chemistry; Serrano et al., 2022) should not be mocked in such a context, not least because they meet certain highly-prized standards of modern science, such as openness and replicability (Frietsch, 2021; Rampling, 2020). In addition, it shows some interesting parallels with modern AI and machine learning techniques in some people's eyes, e.g. "Alchemy was not only a proto-science, but also a 'hyper-science' that overpromised and underdelivered." (Dijkgraaf, 2021, n.p.)

1991), the Soviet Union invited "Bulgaria, Czechoslovakia, the GDR, Hungary, Poland and Romania [...] to join the Unified System of computers" (Szabó, 2020; Cuba joined later; Levin, 2016a, 2016b). This being said, this was not a decision the USSR and the other Comecon countries arrived at lightly[7] (Malinovsky et al., 2010; Пржиялковский, 1997).

> Some were afraid that it was almost impossible [to reverse engineer the System 360] and, for that reason, there was no necessity in direct reproduction of the prototypes' structures. Many suggested to improve the foreign solutions according to designers' own ideas and individual understanding, which naturally were very different.
>
> Vladimir Konstantinovich Levin (2016b, n.p.)

But arrive at it they did, and manufacturing plants were chosen to be inter alia in Minsk, Yerevan, and Moscow (Levin, 2016b). The trail here goes colder, but we know "the Poles had documents on the input/output interface, the East Germans on chip sets, and the Soviets glossaries of terms and algorithms." (Petrov, 2023, p. 87; row *Search strategy*, Table 1)

Importantly, EC ЭВМ clones

> are not a reverse engineering of the IBM s/360 machines in the sense that [they are] exact (or nearly exact) copies[.] That would imply duplication down to the level of circuit components and, if truly successful, interchangeability of parts between the original and the copy. This level of reverse engineering of a major computer system has never been carried out anywhere in the world. [EC ЭВМ] might be described as an effective functional duplication. The architecture, instruction set, and data channel interfaces are the same, permitting the use of IBM software and interchange at the CPU or major subsystem level with relatively little difficulty.
>
> Seymour E. Goodman (1979, p. 556)

Interestingly, the EC ЭВМ's hardware is sometimes superior (Central Intelligence Agency, 1973) or utterly unrelated to the original, showing deep creativity and understanding of what is being reverse engineered, e.g.

> the ES-1040 displays an aggressive design approach in that it has an instruction logic that does not appear to have been taken from any Western machine[.] The ES-1040 has memory interleaving, instruction look-ahead, and an instruction stack for three 64-bit words that permits queuing of up to six instructions. No IBM model had all of these features in a single model.
>
> Central Intelligence Agency (1980, p. 6)

In addition, not only are the internal components divergent to the original IBM machines, the way these systems were built is also different, e.g. "hand assembly" (Central Intelligence Agency, 1989, p. 6) as opposed to presumably a more automated approach by IBM (recall rows *Equivalence sought*, *Multiple realisation*, and *Solution instances*, Table 1).

[7] This reflects similar deliberations and worries within IBM, e.g. the CEO Thomas J. Watson Jr. said: "[the System/360] was the biggest, riskiest decision I ever made, and I agonised about it for weeks, but deep down I believed there was nothing IBM couldn't do." (Watson & Petre, 2000, p. 295)

By 1971, the EC ЭВМ was a reality, with the first mainframe in the series being produced (Davis & Goodman, 1978; Пржиялковский, 1997). So a mere half decade after IBM's original System 360 series started being sold, the USSR and Comecon countries, had reverse engineered it, creating a faithful clone that could run firmware and software written for IBM's hardware (row *Verification method*, Table 1). Perhaps surprisingly, the EC ЭВМ computers are overwhelmingly seen in a negative light (e.g. Malinovsky et al., 2010; Ter-Ghazaryan, 2014) and largely underrepresented in the literature (at least in English language sources; Jankowska, 1993), but this project clearly played a positive role in the dramatic and successful computerisation of the participating countries (Impagliazzo & Proydakov, 2011; Levin, 2016b; Szabó, 2020). For our purposes, lessons from this project are valuable to draw out problems with computationlists' metatheoretical manoeuvres. But first we move to formally treating this true type of reverse engineering.

4.2 Formally searching for computers

Reverse-Engineering Computers

Given: A set of possible electronic elements $\mathcal{E}$, a class of possible computer architecture designs $\mathcal{D}$, and a target computer $\mathbf{t}$ (whose architecture and elements are initially unknown).

Task: Construct a computer $\mathbf{c}$ of type $d \in \mathcal{D}$ from elements in $\mathcal{E}$ such that $\mathbf{c}$ and $\mathbf{t}$ are *functionally equivalent* (denoted $\mathbf{c} \equiv \mathbf{t}$). Here, two computers $\mathbf{c}$ and $\mathbf{t}$ are said to be functionally equivalent if and only if the following two conditions are met:

a) any computer program P that can run on $\mathbf{t}$ can also run on $\mathbf{c}$; and

b) for any input i_p for P, the output of $\mathbf{c}(P, i_p) = \mathbf{t}(P, i_p)$.

Reverse engineering a computer can formally be conceptualized as searching a space $\mathcal{D}$ that contains all hardware architecture designs, and a space $\mathcal{E}$ that contains all sets of electronic elements, for an architecture, $d \in \mathcal{D}$, and a set of elements, $\mathbf{e} = \{e_0\ e_1 \dots e_n\} \in \mathcal{E}$, where $\mathcal{E}$ is the power set of all unique elements. Unlike before, applying $d(\mathbf{e})$ produces a computer $\mathbf{c}$ and not a substance (row *Computation performed*, Table 1). We can exclude from $\mathcal{D}$ designs that map the input to the output directly — and from $\mathcal{E}$ we can exclude any singleton sets since we are looking for how to make $\mathbf{c}$ and not for an electronic device that is ready-made to be equivalent to our target. However, having our target $\mathbf{t}$ or an equivalently detailed functional specification of it handy is imperative for knowing we have in fact found what we are seeking (if not multiple targets: one to remain untouched, many to dissassemble; Rekoff, 1985). Unlike before, correlation is not (straightforwardly at least; more below) a possible metric for knowing if we have reached the target (row *Verification method*, Table 1).

In order to know whether we have found a valid pair $(d, \mathbf{e})$ when $\mathbf{c}$ is produced we need to compare $\mathbf{c}$ to a functional specification, which is both formally defined, if we are lucky exhaustively so, and empirically brute-forced, i.e. various hardware, firmware and software testing (row *Equivalence sought*, Table 1). The first requirement involves making sure that the

system is Turing-complete at some level of engineering abstraction, e.g. the electronics level, by ensuring $(d, \mathbf{e})$ produces NAND gates, flip-flops, latches or some other complete piece of logical hardware components and that these work, e.g. have a power supply, etc.[8] We may also wish to specifically search for registers, CPUs, and so on, but this is not required as other circuit components can also be developed instead.

As a reverse engineer, we are aware of basic computer scientific and hardware engineering principles that are at play to create a general purpose digital computer, and so can also use this knowledge as required (row *Solution instances*, Table 1). This is equivalent to assuming that porcelain, at least as made by the experts in Jingdezhen is something made from minerals (ingredients mined from the ground) fired in a kiln, or assuming that the shape and size of the assembled ingredients one places in the kiln does not change significantly after firing, unlike say dough after baking. Furthermore, this equivalence between the thinking process highlights that in the case of bone china, a type of porcelain, one of the ingredients is bone ash (made from animal bones, which are not mined from the ground; Özgündoğdu, 2005). So just because some ingredients/components or methods/architectures might seem right *a priori*, they are certainly not the only way to get to the target (row *Search strategy*, Table 1). Notwithstanding, in the case of hardware reverse engineering we have very strong reasons to believe, and therefore detect with our expertise and cognition generally, that dissociable modules (for testing, manufacturing, etc.) were used in the design process. And so "[i]f the module boundaries [are used] as the boundaries of the entities for which the specifications are being prepared, the preparation of these specifications can be greatly expedited." (Rekoff, 1985, p. 251)

The second step of the verification process is exactly why we are reverse engineering in the first place: we want firmware and software written by others for the target system to run on our clone. This is carried out once we have decided the system is ready to be tested, i.e. not if we just have a candidate pair $(d, \mathbf{e})$ that produces a system that does not turn on if it uses electricity and if we choose to create an electronic computer, mutatis mutandis for any other way of powering our clone. And so, once we decide it is ready to run our library of programs by running on $\mathbf{c}$ a set of firmware and software, $\mathcal{P} = \{P_1, P_2 \dots P_m\}$, which it can be verified against if and only if we know what behaviour $\mathbf{c}$ is meant to display given $\mathcal{P}$. This can be automated as a series of tests, but failures will not necessarily be informative, only successes. In other words, if the system $\mathbf{c}$ we have built fails to run the programs $\mathcal{P}$ we have in our collection, we cannot infer *how* we need to fix it, what changes to make or what went wrong necessarily. We can only know *that* we need to fix it, we must make some kind of modification to obtain a better candidate pair $(d, \mathbf{e})$.[9]

Industrial espionage and obtaining legal licences play a role, as before in the case of porcelain, in trimming the search space down significantly (Central Intelligence Agency, 1973; Goodman, 1979; Levin, 2016a; Malinovsky et al., 2010; Petrov, 2023; Szabó, 2020). In the case of the EC ЭВМ, knowledge of IBM's Extended Binary Coded Decimal Interchange Code (EBCDIC) character encoding and of the assembly language and corresponding machine code — i.e. the finite set of all commands (and what they do) of which all programs are made of — is enough to prove that applying $d(\mathbf{e})$ gives us a $\mathbf{c}$ we are satisfied with. In other words, having access to documents such as Fagg et al. (1964) provides serious hints as to what to build and how. We need not brute force test using the library $\mathcal{P}$ if we have the machine code, except for checking that the hardware peripheral input/output devices work and that the time and space complexity of the hardware is sufficient.

Based on the two verification steps above, we can infer: the first is tractable, in the sense that making Turing-complete components is a process that is well-known to us as hardware (reverse) engineers. The second step is incredibly difficult, if not entirely intractable, without peeking at the methods the original engineers used (Rich et al., 2021; van Rooij et al., 2024). Nothing other than knowing what we want the output to look like is guiding us, which provides no guidance within a formal framework — without knowing how IBM built their mainframes, only informal intuition and luck can stand a chance at aiding us (recall this took less than a decade; row *Search duration*, Table 1). On this point, others have commented on the reversing accomplished by those working on the EC ЭВМ.

> As more and more circuits are etched into a smaller and smaller space, the task of copying or reverse engineering a microprocessor becomes increasingly difficult for the person without knowledge of the designers' original intent.
>
> David A Wellman (1989, p. 80)

What perhaps causes confusion or a loss of intuition is that on the one hand there appear to be infinite possible instantiations of general purpose computers (rows *Multiple realisation* and *Solution instances*, Table 1), while on the other hand (as we shall see below) no guarantee is given to find such a system because the search space is even larger than the space of plausible candidate solutions. In the cognitive science setting, no peeking at the solutions, or at hints, can help us shrink the space. No industrial espionage is possible. No intent has engineered cognition (cf. Hardcastle & Hardcastle, 2015; Lee & Johnson-Laird, 2013; Rekoff, 1985; Tennor, 2015; Vaesen & van Amerongen, 2008).[10] Brains and minds are not engineered systems and so inferences that apply in reverse engineering proper fall apart in a cognitive scientific context. These confusions — which manifest as shuttling between the two incompatible columns in Table 2 — are what we will unpack below.

5 Mind-as-substance versus mind-as-computer

> A network of commitments is in reflective equilibrium when each of its elements is reasonable in light

[8] The mere existence of say NAND gates in our collection of electronic components $\mathcal{E}$ does not ensure Turing-completeness in $\mathbf{c}$. They are not sufficient for a machine to display completeness, we need to make sure $\mathbf{c}$ is Turing-equivalent.

[9] In general, determining whether two computers $\mathbf{c}$ and $\mathbf{t}$ are functionally equivalent is uncomputable (Rice, 1953), i.e. there cannot exist any computational procedure guaranteed to solve this problem (row *Verifiability* in Table 2).

[10] In the framework of those who propose evolution has some teleological intent-like properties (e.g. Boudry & Pigliucci, 2013; Csete & Doyle, 2002; Dennett, 1995; Dretske, 2000; Erneling & Johnson, 2005; Figdor, 2023; Harnad, 2003; Jonas & Kording, 2017; Leahey, 2005; Simon, 1995), there seems no sensible way to discern it either.

Table 2: Comparison of the differences between viewing the mind as a substance and viewing it as a computer under reverse engineering as a neurocognitive methodology. The columns represent the two theses' mutually exclusive properties and vocabularies, e.g. in row *1*) what type of analysis is carried out: mechanistic (column *Property*, **Mind-as-substance**) versus functional (column *Property*, **Mind-as-computer**); and what language used is consistent with each approach, e.g. the phrase 'interactions amongst components' (in pink, column *Vocabulary*, **Mind-as-substance**) versus 'input-output mapping' (in grey, column *Vocabulary*, **Mind-as-computer**). The final column, **Mix-and-match** represents the metatheoretical calculus, the pattern we see in the literature, which is what it says on the tin: a mishmash of the previous two theses that computationalists are forced to contort their science into. In the main text, extracts contain highlighted phrases using the colour matching that of the respective column in the table above. Importantly, it is not correct usage to use single words or phrases in the *Vocabulary* columns to diagnose what thesis, mind-as-substance or -as-computer, is being used when we notice the reverse engineering methodology being deployed. For example, we do not propose that seeking a mechanistic understanding is necessarily equivalent to the thesis mind-as-substance without minimally a commitment to reverse engineering, makeism, and/or computationalism being established, and even then one must exercise caution (Guest & Martin, 2021).

Neurocognitive Thesis Under Reverse Engineering as Methodology					
	Mind-as-substance		**Mind-as-computer**		**Mix-and-match**
	Property	*Vocabulary*	*Property*	*Vocabulary*	*Property*
1) *Analysis*	mechanistic	interactions amongst components, machinery, mechanism, similar to biological neurons, structure	functional	abstract, behaviour, cognition, computation, function, input-output mapping, task	both in name, but neither causal mechanistic explanation nor functional decomposition is provided
2) *Correlation with data*	required, informative	benchmark, data, dataset, correlation, measure, predict	irrelevant, misleading	cognition, computation	correlation is deemed required and informative
3) *Multiple realisability*	rejected	correlation, similar to biological neurons, structure	in full swing	abstract, algorithm, cognition, computation	implicit and explicit shuttling between rejection and acceptance
4) *Number of instances*	one or few	anatomical, structure	infinite	algorithm, cognition, computation	shuttling between uniqueness of mind and remaining unaddressed
5) *Structural similarity*	required	anatomical, interactions amongst components, machinery, measure, mechanism, similar to biological neurons, structure	irrelevant, impossible	abstract, cognition, computation	shuttling between required, e.g. models are brain-like, and irrelevant, e.g. models are abstract
6) *Verifiability*	possible	correlation, benchmark, data, dataset, measure, predict	uncomputable	algorithm, cognition, computation	datasets, benchmarks, and correlations deemed arbiters

> of the others, and the network as a whole is as reasonable as any available alternative in light of our relevant previous commitments. Even if some components would be doubtful in isolation, collectively they constitute an interwoven tapestry of commitments that we can on reflection endorse.
>
> Catherine Z. Elgin (2017, p. 4)

What do we learn about reasoning in cognitive (neuro)science from the above analysis of reverse engineering a substance versus a computer? Through the side-by-side contrast we see in Table 1 we can infer, based on the method a scientist uses, what must follow about their beliefs and vice versa, culminating in the conclusion presented in Table 2 about what their thesis, their "network of commitments", about the nature of cognition could be. Furthermore, it is possible we can evaluate "how the network of commitments hangs together when we recognize how it might fall apart, how easily it might fall apart, and what the consequences of its doing so would be." (Elgin, 2017, p. 307). The analysis captured by column **Mix-and-match**, Table 2, comprises the metatheoretical calculus of the field with respect to "reverse engineering" *as conceived by cognitive neuroscientists* (Guest, 2024; Guest & Martin, 2023,

2024): taking a little from column A, a little from column B. The problem is that the two columns are incompatible. And so we encounter problematic methodological manoeuvres both in principle and in practice. In the remainder of this section, not only do we see self-described computationalists methodologically treat mind as a substance — a position incompatible with core computationalist axioms — we also see a mixing and matching of methodologically treating mind as a substance and as a computer in violation of both of these mutually exclusive belief systems (column **Mix-and-match**, Table 2).

Foundational cases of methodological proposals have already been touched on in section 2, **Cognitive science as reverse engineering**, where we can see in light of Table 2 how the commitments by philosophers of mind and of cognitive science are themselves at odds with seeing mind either as a computer or as a substance, instead framing them as a mishmash of both, which is in violation both of consistency and of computationalism.[11] Using each row in Table 2, we can discern that practitioners:

1) attempt both a functional and a mechanistic analysis in name, but avoid a functional decomposition and/or an analysis of the causal mechanistic components in practice;

2) prize correlation with data, where the more correlations between computational model and neuroimaging and behavioural data, the better;

3) reject multiple realisability, implicitly or accidentally, especially when explicitly stating there are few candidate instances of minds;

4) believe that we are somehow close to (reverse) engineering brain, cognition, and/or behaviour instead of acknowledging there are infinitely many candidate solutions under computationalism;

5) seek similarity between putative structures in the brain and their models; and finally,

6) assert that more and more data collection is a necessary activity to evaluate, or furthermore validate as brain-like, our computational models.

Such methodological properties place scientists predominantly in the mind-as-substance column, something perhaps discombobulating and disparaging to contend with if one is a computationalist (captured in column **Mix-and-match**, Table 2). To further elucidate this, in Table 2 words and phrases have been highlighted in colour as a function of which thesis they are consistent with, see *Vocabulary* columns under **Mind-as-substance** (pink) and **Mind-as-computer** (grey) columns. For example:

> In our framing, reverse engineering consists of figuring out the input-output mapping for all neurons and muscle cells as well as the inputs from the world (i.e. system identification), then reassembling the collection of input-output functions into a robust, simulatable model that can exhibit key behaviors when connected to the simulated body. To be successful, this working model should recapitulate behavior under a range of conditions, stimuli, and perturbations.
>
> Gal Haspel et al. (2023, reference to figure removed, n.p.)

Another example of this pattern of reasoning is present here:

> Recently, a new "reverse engineering" approach to computational modeling in systems neuroscience has transformed our algorithmic understanding of the primate ventral visual stream (Bao et al., 2020; Cadena et al., 2019; Cichy et al., 2016; Kietzmann et al., 2019; Kubilius et al., 2019; Schrimpf et al., 2018, 2020; Yamins et al., 2014) and holds great promise for other aspects of brain function. This approach has been enabled by a breakthrough in artificial intelligence (AI): the engineering of artificial neural network (ANN) systems that perform core perceptual tasks with unprecedented accuracy, approaching human levels, and that do so using computational machinery that is abstractly similar to biological neurons. [ANNs] for object recognition have now been shown to predict the responses of neural populations in multiple stages of the ventral stream (V1, V2, V4, and inferior temporal [IT] cortex), in both macaque and human brains, approaching the noise ceiling of the data.
>
> Martin Schrimpf et al. (2021, citation style modified, p. 1)

Explicit commitments to mind-as-computer and to reverse engineering as methodology are made often (e.g. DiCarlo, 2018; Dupoux, 2018; Haspel et al., 2023; Jonas & Kording, 2017; Schrimpf et al., 2021; Zednik, 2018). These examples explain, in other words, that they perform an analysis that is functional (row *Analysis*, Table 2) and black box with respect to the system they are interested in, e.g. "solving the problem by machine learning from large datasets" (Haspel et al., 2023, n.p.) and

> an alternative approach 'reverse-engineering' (DiCarlo, 2018): effectively searching through the very large class of all possible neural network models by iteratively improving the current best model, based on guidance from benchmarks.
>
> Martin Schrimpf et al. (2020, p. 420)

However, the reliance on benchmarks and datasets places them in mind-as-substance (rows *Correlation with data*, *Number of instances*, and *Verifiability*, Table 2). In other words, these cognitive (neuro)scienists want an "input-output mapping" (Haspel et al., 2023, n.p.) provided by the ANN, essentially used as a statistical model since it is used without any deeper functional decomposition into cognitive capacities (and this is quasi-scientific at best, viz. Guest & Martin, 2024; Singmann et al., 2022; Sullivan, 2022). The models perform something more akin to function approximation, both in practice and in their own words (Guest & Martin, 2024). This strategy is distinct from brute forcing a functional specification of the target system, since "implementations are not specifications" (Cooper & Guest, 2014). This is an even lower bar than that sought in section 4, **Reverse engineering computers**.[12] In

[11] The authors condone neither a mixing of the two nor either individually: reverse engineering is not a viable scientific methodology (see section 6, **Will the real computationalist please stand up?**).

[12] But note that use of benchmarks, and of functional approximation to them,

many such examples, there is further elaboration: "The core of reverse engineering a nervous system is figuring out how interactions among components (here neurons) shape neural dynamics and behavior" (Haspel et al., 2023, n.p.). As well as:

> Only synthesis in a computer simulation can reveal what the interaction of the proposed component mechanisms actually entails and whether it can account for the cognitive function in question. If we did have a full understanding of an information-processing mechanism, then we should be able to engineer it.
>
> Nikolaus Kriegeskorte and Pamela K. Douglas (2018, p. 1148)

This shifts back to a mechanistic analysis, given their focus on proposed causal interactions between the theorised relevant components (typical of neuroscience; Chirimuuta, 2013; Guest & Martin, 2025; Miłkowski, 2016; Ross & Bassett, 2024; Zednik, 2014). These cases are explicit, e.g. "integrative benchmarking [is] the next step to building neurally mechanistic models of domains of human visual intelligence."(Schrimpf et al., 2020, p. 420) Looking at Table 2's row *Analysis*, when previously we were set up to expect a functional analysis to match computationalism, we are now firmly placed in mind-as-substance territory.

Another important aspect are claims like: "If we had a way of hypothesizing the right structure, then it would be reasonably easy to test." (Jonas & Kording, 2017, p. 16) Also:

> in silico experiments would then allow us to build understanding: to interpret the dynamics in terms of computational concepts, from decision-making, memory, and sensory integration to attention and coordination, and indeed to understand fundamental principles of circuit structure and function.
>
> Gal Haspel et al. (2023, n.p.)

On the one hand, the analysis is one of a cognitive (e.g. reference to decision-making, attention, etc.) and computational system (row *Analysis*, Table 2). However, on the other hand, they ask:

> Can we easily be misled and believe we understand how a nervous system works from partial recording? How probable is it that the models we fit get the correlations right and the causality wrong (Tremblay et al. 2022)? How much data of what kind is too little to reverse engineer systems?
>
> Gal Haspel et al. (2023, n.p.)

Here, they instead underline that using correlations as evaluations of (statistical) model fit is potentially misleading (row *Correlation with data*, Table 2).

Correlations are utterly misleading if one subscribes to the mind-as-computer thesis. Two computers can be 100% identical under any meaningful definition of identical, e.g. two laptops can have identical hardware specifications, but have utterly uncorrelated pixels on their screens as well as contents loaded in their memory or stored on their hard disk drives, e.g. somebody can be watching a video on full screen, and another user can have various windows open to do with programming. And the opposite is true, a Raspberry Pi (a computer that is built on a single circuit board) and a typical laptop can have identical videos playing, but their hardware is divergent. In the final question quoted, they invoke the amassing of data, stating the more, the better. In this context, this is irrelevant (both philosophically and statistically, e.g. Davis-Stober & Regenwetter, 2019; Douglas, 2000; Duhem et al., 1982; Egan, 1999; Elgin, 2017; Kellen et al., 2021; Lasonen-Aarnio & Littlejohn, 2024; Longino, 1990; Quine, 1953) and the attention to such aspects is incompatible with reverse engineering computers (rows *Correlation with data* and *Verifiability*, Table 2). These tensions appear unacknowledged, with frequent shuttling between mind-as-substance and -as-computer:

> Brains are the product of evolution and development, processes that are not constrained to generate systems whose behavior can be perfectly captured at some abstract level of description. It may therefore not be possible to understand cognition without considering its implementation in the brain or, conversely, to make sense of neuronal circuits except in the context of the cognitive functions they support.
>
> Nikolaus Kriegeskorte and Pamela K. Douglas (2018, p. 1157)

Brains can only be understood "abstractly" (i.e. functionally) if we seek computational understanding.

Further examples of toing and froing can be seen in statements like: "We are not only interested in having correct model outputs (behaviors) but also internals that match the brain's anatomical and functional constraints" (Kubilius et al., 2019, p. 3); and:

> Understanding the computational mechanisms of human vision therefore requires us to measure and model the rapid representational dynamics across the different regions of the ventral stream.
>
> Tim C. Kietzmann et al. (2019, p. 21854)

These two positions are irreconcilable in light of Table 2, especially rows *Multiple realisability*, *Analysis*, and *Structural similarity*.

The principle of multiple realisability (Chirimuuta, 2018, 2021; Egan, 2017; Figdor, 2010; Guest & Martin, 2023; Hardcastle, 1995, 1996; Litch, 1997; Polger & Shapiro, 2016; Ross, 2020) in the context of mind-as-computer causes a cascade of further multiple realisability:

> computational states are doubly multiply realizable. In addition to being realizable within a multitude of disparate objects, a computational state is also realizable in any of an infinite number of (continuously varying) physical states within the same object.
>
> Mary Litch (1997, p. 359)

Therefore, it becomes clear that we have no hope of finding any single "computational state" (Litch, 1997). And this is the case even if we can peek at the engineered system's specification, because computational states are not important for the reverse engineering of computers. Only functionalism stands

is only reasonable as a reverse engineering method if multiple realisability is rejected. Since, in the face of multiple realisability that follows from mind-as-computer there is massive underdetermination of algorithms and implementations for any given (approximate) function.

a chance of providing us with needed answers in this scheme (Chirimuuta, 2018; Egan, 2017; Figdor, 2010; Guest and Martin, 2023; Hardcastle, 1995; McCarthy and Keil, 2023; van Rooij and Baggio, 2021, cf. Miłkowski, 2016).

> As Sejnowski puts it, "Although a working model can help generate hypotheses, and rule some out, it cannot prove that the brain necessarily solves the problem in the same way" (Sejnowski et al., 1988, p. 1304). In other words, simulating the behavior only shows that you have a candidate explanation; it does not show that you have the right explanation, i.e. one that produces the behavior in the "same way".
>
> Catherine Stinson (2018, p. 126)

Ways of deciding to consult Table 2, can be noticing when a research programme is described as "reverse engineering the human cognitive toolkit" (Fan, 2020, n.p.) or as "seek[ing] to 'reverse engineer' [the brain's] algorithms [in order] to learn both about how our minds work and build more effective artificial intelligence systems" (Yamins, 2019, n.p.) or as: "Reverse engineering the mind, or understanding the computational principles that give rise to cognition, is a common goal of cognitive science, artificial intelligence, and neuroscience." (Groen, 2022, n.p.) Although, the methodology and metatheoretical reasoning we critique need not be labelled as 'reverse engineering' by practitioners.[13] And so the burden is, unfairly perhaps, on the reader to notice its deployment through the prism of Table 2.

A metatheoretical calculus that consists of proposing reversing the mind and shuttling between the mind-as-substance and -as-computer theses should be troubling (column **Mix-and-match**, Table 2). A scholar who reasons the ways we have metatheoretically mapped herein may not be a computationalist. If they are indeed not computationalist, now is the time to rid themselves of the computationalist and functionalist vocabulary, constraints, and concepts. However, if they are computationalist, then our analysis here should be a wake-up call. A scholar who subscribes to makeism, i.e. intends to stay firmly within the mind-as-computer thesis as presented in Table 2, also needs to perhaps reconsider their makeist stance,[14] but regardless needs to take heed of the incompatibility of their views in light of their methodological practice. On the contrary, if one is to stay within the bounds of the mind-as-substance thesis as a computationalist, they should be cautious, not least because it gives rise to a category mistake (viz. Egan, 2020; Rozemond, 1998; Ryle, 2009/1949; Sayward, 1983). These issues are all the result of methodological and metatheoretical manoeuvres performed to avoid the "falling apart" of the "network of commitments", but instead constitute its disintegration as such (recall quote from Elgin, 2017). In other words:

a) reverse engineering coupled with the mind-as-computer thesis falls under makeism with its many traps (column **Mind-as-computer**, Table 2; van Rooij et al., 2024);

b) reverse engineering coupled with the mind-as-substance thesis is prima facie flawed when deployed by computationalists because it is incompatible with computationalism (column **Mind-as-substance**, Table 2); and

c) reverse engineering shuttling between both is a network of commitments in full-blown decay in multiple ways, as outlined above (column **Mix-and-match**, Table 2).

These tensions are core to why reverse engineering should have been dismissed out of hand, along with makeism generally, as well as formal realism (Chirimuuta, 2021), and the dehumanisation and pseudoscience that comes from AI and machine learning, and other limiting, reductionist, and positivist views (Andrews et al., 2024; Birhane, 2022; Birhane & Guest, 2021; Erscoi et al., 2023; Gebru & Torres, 2024; van der Gun & Guest, 2024).

> In conclusion, it is perhaps worth while giving a warning against a very popular fallacy. The hearsay knowledge that everything in Nature is subject to mechanical laws often tempts people to say that Nature is either one big machine, or else a conglomeration of machines. But in fact there are very few machines in Nature.
>
> Gilbert Ryle (2009/1949, p. 68)

6 Will the real computationalist please stand up?

> Oftentimes, however, philosophers of mind and others writing on the philosophical implications of artificial intelligence fail to understand what computation is, and what computation (in the context of the computational model of the mind) implies about mental processes.
>
> Mary Litch (1997, pps. 357–358)

We should be on guard and strive to avoid "ill-posed argumentation [being] unwittingly permitted during (meta)scientific inference" (Guest & Martin, 2023, p. 218). This paper explains the serious reasoning problems that arise when cognitive scientists deploy (knowingly or not) reverse engineering as a method to understand brain, behaviour, and cognition. We explain this using a metatheoretical calculus, embodied in Table 2. We demonstrate what it means when we as cognitive scientists insist on using reverse engineering as a methodology. Unbeknownst to us, our beliefs become contorted in grotesque ways, when we appear to shuttle between the two theses mind-as-substance versus -as-computer (cf. Hardcastle, 1995; Leibniz, 1714; Rozemond, 2014; Ryle, 2009/1949; Sayward, 1983). In other words, we show that the methodological commitments and forms of reasoning taken by practitioners do not follow from computationalism, but in fact oppose it (also see Guest & Martin, 2025, Box 1). This is facilitated by and facilitates framing the mind as a substance and not as a system that performs computations. We show that mutually exclusive aspects are held to be possible, e.g. scientists seek structural similarity while also proclaiming to support the idea of mind-as-computer. But the models, e.g. ANNs, are nonetheless run

[13] In fact, the examples given in this paragraph (Fan, 2020; Groen, 2022; Yamins, 2019) are not taken from archival sources, but scientists' writings outside the official literature.

[14] For instance, it is known that "[s]imulation of the brain by a computer is unlikely not because of the massive computational power of the brain, but because of the overhead required when one model of computation is simulated by another" (Parberry, 2013, p. 125). In other words, if one wants hold on to computationalism and wants to reverse engineer cognition then makeism seems more feasible if one does not even try to simulate the brain at the implementational level. It also is not necessary, since from the mind-as-computer perspective only *functional equivalence* is sought.

on a substrate that is a von Neumann machine, which should be anathema to those who think structural similarity is important, and to those with a mind-as-substance view (recall section 5, **Mind-as-substance versus mind-as-computer**). The same goes for the black box, i.e. functional analysis approach (Petrick, 2020) — how can a model we do not understand furnish us with mechanistic understanding? That is to say, statements by practitioners are made that contradict their own commitments to computationalism, to multiple realisability, and to what functional analysis can grant us even in the best case, e.g. "the problem of reverse engineering a computational system, including the human mind, seems to inevitably move primarily from function to mechanism." (Chater et al., 2011, p. 196) This move simply cannot be the case.

These reasoning issues, we posit, stem from philosophers of science, and especially of mind and of brain sciences, asserting makeist claims such as Dretske (1994)'s titular claim *If You Can't Make One, You Don't Know How It Works* (e.g. Dennett, 1995; Kriegeskorte and Douglas, 2018; Miłkowski, 2013; for critique and history see: van Rooij et al., 2024). These views themselves have their origins in the cybernetics movement of the 40s and 50s, which first employed engineering and other formal methodologies in "transformative and imperialist" ways to the brain and psychological sciences broadly construed (Abraham, 2012; Petrick, 2020). The type of makeism we tackle herein takes the form 'if we cannot make x, then we cannot understand x'. Through contraposition this provides us with 'if we understand x, then we can make x', where x is brain, cognition, and behaviour. Such syllogisms if taken at face value mean we will never know how the mind or brain "work" because, under computationalism, we cannot realistically, i.e. tractably, make such a system in practice (Rich et al., 2021; van Rooij et al., 2024). So when proponents argue that, e.g. "[t]he history of psychology suggests that well-specified task analyses (Marr, 1982) are the most tractable way of reverse engineering the structure of cognitive mechanisms" (Pietraszewski & Wertz, 2011, p. 209), they can be safely rejected out of hand. Rich et al. (2021) and van Rooij et al. (2024) explain, using formal proofs, why tractability is not guaranteed by any method, scientific or engineering. Ultimately, "the brain is a [...] non-decomposable system, [and so] reverse engineering [as a methodology] must necessarily fail and will not provide the envisaged understanding!" (Schierwagen, 2012, p. 149)

An important caveat is that some scholars use this phase metaphorically (Polger & Shapiro, 2023). For example, Natalia Levshina (2021) explains that she uses it as an analogy: "Similar to reverse engineering, we determine which of the probabilistic measures could have been responsible for the recurrent cross-linguistic patterns described in the literature." (p. 1) Also, Iris van Rooij and Todd Wareham (2008) use reverse engineering as a contrast for the type of (forward) engineering that pervades computer science as an engineering field, i.e. "cognitive science is engaged in a form of reverse engineering, to be contrasted with the forward engineering that typically occurs in computer science" (p. 2). Nina Poth (2022) also outlines that our own set of stances (Guest & Martin, 2021; van Rooij & Baggio, 2021) "resonates well with the reverse engineering perspective" (n.p). To such uses of language, since no formal reversing method is described, we say that increasing care should be taken when deploying the phrase "reverse engineering" metaphorically or otherwise, as its baggage is becoming increasingly heavy, with non-makeist stances perhaps requiring adjustment into actively anti-makeist ones (Guest & Martin, 2023; van Rooij et al., 2024). Computationalism need not be makeist — it need not subscribe to reverse engineering (Marom, 2009; Rueckl, 2012; Schierwagen, 2012) nor to implementational multiple realisability (Miłkowski, 2016; Polger and Shapiro, 2023 as a practical reality; but recall Litch, 1997, on double multiple realisability above).

To recapitulate, no examples (can) exist of literally reverse engineering non-human-engineered systems — only of criticisms of attempting to do so (e.g. Marom, 2009; Rueckl, 2012; Ryle, 2009/1949; Schierwagen, 2012). "[Cognition] cannot be taken to bits and the laws it obeys are not those known to ordinary engineers." (Ryle, 2009/1949, p. 9–10) Those who argue for reverse engineering non-engineered systems would be on more solid ground if they accepted what we need are scientific theories and models. What is actually going on is that the "original system" (Schierwagen, 2012), the phenomenon, a cognitive capacity, what have you, has been theoretically analysed, and a model is built of this (formal) specification; this is typical computational modelling of a theory (e.g. Guest & Martin, 2021; van Rooij & Baggio, 2021). There is no ground truth of the original system, any more than there is a ground truth for anything under pessimistic meta-induction (Laudan, 1981). On the contrary, a scientific model is built within a theory and framework, and there is no remaking of cognition (as others also explain: Guest & Martin, 2023; Schierwagen, 2012; Stinson, 2018; van Rooij et al., 2024).

Taken all together, reverse engineering provides unstable footing for the cognitive scientist when thinking about what cognition — mind, brain, behaviour — is, causing erratic picking and mixing of views. To understand our own commitments, metatheoretical and methodological, with respect to cognition, looking at Table 2 suffices to alert us, e.g. if we work within computationalism but get sidetracked into seeking structural similarity. Makeist views without explicit use of reverse engineering do not absolve us from entertaining — accidentally or otherwise — framings that do not obtain. Makeism, reverse engineering, and other related hubristic and misplaced cognitive scientific positions and endeavours, are "why, historically, the use of AI to understand cognition fizzled out, and why it will do so again if we do not change our present course." (van Rooij et al., 2024, p. 625) Much like how alchemy transitioned through chymistry[6] to modern chemistry, by shedding unscientific methods and goals (viz. Frietsch, 2021; Rampling, 2020; Serrano et al., 2022), cognitive science can also abandon imperfect and impossible methodologies, e.g. reverse engineering, and goals, e.g. building makeist models. Exactly because "uncultivated science can easily turn into occult science or into the cult of science" (Stengers, 2018, p. 10) we need to be on high alert.

We must strive to deeply understand that "[humans] are not machines, not even ghost-ridden machines." (Ryle, 2009/1949, p. 67) If it is true that "we can think of engineering as science for people who are impatient" (Simon, 1995, p. 100), then perhaps that is the cause for much of this (e.g. Stengers, 2018). Relatedly, "the engineer's schooling and workshop experience teach [them] to design bridges and not, save *per accidens*, to build or expound theories." (Ryle, 2009/1949, p. 289) The slow, hard, often painful, but never tedious — except for those driven by other motives (Harris, 2023; Legg et al., 2021) — process of doing science is just that. It cannot be sped up, cannot be replaced with (reverse) engineering, cannot be outsourced to AI.

There are no shortcuts. Those who do not take heed of these warnings are doomed to fractally if not frantically mix-and-match from incompatible ideologies. Computationalism may or may not be abandoned, but either way it is not compatible with anything goes.

> Thou shalt not make a machine to counterfeit a *human* mind.
>
> Frank Herbert (1965, p. 18)

References

Abraham, T. H. (2012). Transcending disciplines: Scientific styles in studies of the brain in mid-twentieth century america. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(2), 552–568.

Amdahl, G. M., Blaauw, G. A., & Brooks, F. P. (1964). Architecture of the IBM System/360. *IBM Journal of Research and Development*, *8*(2), 87–101.

Andrews, M., Smart, A., & Birhane, A. (2024). The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, 101027.

Aplin, T. (2013). Reverse engineering and commercial secrets. *Current Legal Problems*, *66*(1), 341–377.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

Birhane, A. (2022). Automating ambiguity: Challenges and pitfalls of artificial intelligence. *arXiv preprint arXiv:2206.04179*.

Birhane, A., & Guest, O. (2021). Towards decolonising computational sciences. *Kvinder, Køn & Forskning*, (2), 60–73.

Boudry, M., & Pigliucci, M. (2013). The mismeasure of machine: Synthetic biology and the trouble with engineering metaphors. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 660–668.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Cauwenberghs, G. (2013). Reverse engineering the cognitive brain. *Proceedings of the National Academy of Sciences*, *110*(39), 15512–15513.

Central Intelligence Agency. (1973). Soviet RYAN Computer Program. In *Foia collection*. https://www.cia.gov/readingroom/document/0000309585

Central Intelligence Agency. (1980). Use of western technology in the Ryad computers of the USSR and Eastern Europe. In *General cia records*. https://www.cia.gov/readingroom/document/cia-rdp85t00176r000900010001-6

Central Intelligence Agency. (1989). Soviet bloc computers: direct descendants of Western technology. In *The princeton collection*. https://www.cia.gov/readingroom/document/0000500644

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*(1), 36–67.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about bayesian cognitive science. *Behavioral and Brain Sciences*, *34*(4), 194–196.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science*, *1*(6), 811–823.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, *10*(7), 287–291.

Chirimuuta, M. (2013). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153.

Chirimuuta, M. (2018). Marr, Mayr, and MR: What functionalism should now be about. *Philosophical Psychology*, *31*(3), 403–418.

Chirimuuta, M. (2021). Your brain is like a computer: Function, analogy, simplification. In F. Calzavarini & M. Viola (Eds.), *Neural mechanisms: New challenges in the philosophy of neuroscience* (pp. 235–261). Springer.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1).

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42–49.

Corporation, R. (1965). Rca spectra 70. https://www.computerhistory.org/brochures/doc-4372956eb9810/

Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, *295*(5560), 1664–1669.

Darden, L. (2007). Mechanisms and models. In D. L. Hull & M. Ruse (Eds.), *The cambridge companion to the philosophy of biology* (pp. 139–159). Cambridge University Press.

Darden, L. (2008). Thinking again about biological mechanisms. *Philosophy of science*, *75*(5), 958–969.

Davis, N. C., & Goodman, S. E. (1978). The Soviet Bloc's Unified System of Computers. *ACM Computing Surveys*, *10*(2), 93–122.

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review*, *126*(6), 865–879.

Dawson, M. R. (2013). *Mind, body, world: Foundations of cognitive science*. Athabasca University Press.

Denić, M., & Szymanik, J. (2022). Reverse-engineering the language of thought: A new approach. *Proceedings of the annual meeting of the cognitive science society*, *44*(44).

Dennett, D. C. (1995). Cognitive science as reverse engineering several meanings of "top-down" and "bottom-up". In *Studies in logic and the foundations of mathematics* (pp. 679–689, Vol. 134). Elsevier.

d'Entrecolles, F. X. (1712). The First Letter from Père d'Entrecolles Missionary of the Order of Jesuite to Father Orry. https://gotheborg.com/letters/letters_first.shtml

d'Entrecolles, F. X. (1722). The Second Letter from Père d'Entrecolles Missionary of the Order of Jesuite to Father Orry. https://gotheborg.com/letters/letters_second.shtml

DiCarlo, J. J. (2018). To advance artificial intelligence, reverse-engineer the brain. *Wired magazine Opinion*. https://www.wired.com/story/to-advance-artificial-intelligence-reverse-engineer-the-brain

Dijkgraaf, R. (2021, October). The uselessness of useful knowledge. https://www.quantamagazine.org/science-has-entered-a-new-era-of-alchemy-good-20211020/

Donig, S. (2010). Appropriating American technology in the 1960s: Cold War politics and the GDR computer industry. *IEEE Annals of the History of Computing*, *32*(2), 32–45.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579.

Dretske, F. I. (1994). If you can't make one, you don't know how it works. *Midwest studies in philosophy*, *19*, 468–482.

Dretske, F. I. (2000). *Perception, knowledge and belief: Selected essays*. Cambridge University Press.

Duhem, P., Wiener, P. P., & Vuillemin, J. (1982). *The aim and structure of physical theory* (Vol. 126). Princeton University Press. Retrieved August 16, 2024, from http://www.jstor.org/stable/j.ctv1nj34vm

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Egan, F. (1999). In defence of narrow mindedness. *Mind & Language*, *14*(2), 177–194.

Egan, F. (2017). Function-theoretic explanation. *Explanation and integration in mind and brain science*, 145–163.

Egan, F. (2020). A deflationary account of mental representation. In *What are mental representations* (pp. 26–53). Oxford University Press New York.

Eilam, E. (2011). *Reversing: Secrets of reverse engineering*. John Wiley & Sons.

Elgin, C. Z. (2017). *True enough*. MIT press.

Encyclopædia Britannica. (2024). Porcelain. https://www.britannica.com/art/porcelain

Erneling, C. E., & Johnson, D. M. (Eds.). (2005). *The mind as a scientific object: Between brain and culture*. Oxford University Press.

Erscoi, L., Kleinherenbrink, A., & Guest, O. (2023). Pygmalion displacement: When humanising AI dehumanises women. *SocArXiv*. https://doi.org/10.31235/osf.io/jqxb6

Fagg, P., Brown, J., Hipp, J., Doody, D., Fairclough, J., & Greene, J. (1964). IBM System/360 engineering. *Proceedings of the October 27-29, 1964, fall joint computer conference, part I*, 205–231.

Fan, J. (2020). About the lab. https://cogtoolslab.github.io/

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science*, *77*(3), 419–456.

Figdor, C. (2023). What are we talking about when we talk about cognition?: Human, cybernetic, and phylogenetic conceptual schemes. *The Journal for the Philosophy of Language, Mind and the Arts*.

Finlay, R. (2010). *The pilgrim art: Cultures of porcelain in world history*. University of California Press.

Frietsch, U. (2021). Alchemy and the early modern university: An introduction.

Fyrbiak, M., Strauß, S., Kison, C., Wallat, S., Elson, M., Rummel, N., & Paar, C. (2017). Hardware reverse engineering: Overview and open challenges. *2017 IEEE 2nd International Verification and Security Workshop (IVSW)*, 88–94.

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, *29*(4).

Gillette, M. B. (2016). *China's porcelain capital: The rise, fall and reinvention of ceramics in Jingdezhen*. Bloomsbury Publishing.

Goodman, S. E. (1979). Soviet computing and technology transfer: An overview. *World Politics*, *31*(4), 539–570.

Gray, G., & Smith, R. (2001). Sperry Rand's third-generation computers 1964–1980. *IEEE Annals of the History of Computing*, *23*(1), 3–16.

Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: Reverse engineering the mind*. MIT Press.

Groen, I. (2022). Introduction to computational cognitive neuroscience. https://datanose.nl/Course/Manual/110584/Introduction%20to%20Computational%20Cognitive%20Neuroscience/2022

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, *7*(4), 508–522.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*(2), 213–227.

Guest, O., & Martin, A. E. (2024). A metatheory of classical and modern connectionism. *PsyArXiv*. https://osf.io/preprints/psyarxiv/eaf2z_v1

Guest, O., & Martin, A. E. (2025). Are neurocognitive representations 'small cakes'? https://philsci-archive.pitt.edu/24834/

Gurney, K. N. (2009). Reverse engineering the vertebrate brain: Methodological principles for a biologically grounded programme of cognitive modelling. *Cognitive Computation*, *1*, 29–41.

Gutiérrez, A., Gerrard, C., Zhang, R., & Guangyao, W. (2021). The earliest Chinese ceramics in Europe? *Antiquity*, *95*(383), 1213–1230.

Gyekye, K. (1973). An examination of the bundle-theory of substance. *Philosophy and Phenomenological Research*, *34*(1), 51–61.

Hardcastle, V. G. (1995). Computationalism. *Synthese*, *105*, 303–317.

Hardcastle, V. G. (1996). *How to build a theory in cognitive science*. State University of New York Press.

Hardcastle, V. G., & Hardcastle, K. (2015). Marr's levels revisited: Understanding how brains break. *Topics in Cognitive Science*, *7*(2), 259–273.

Harnad, S. (2003). Minds, Machines and Turing. *The Turing Test*, 253–273.

Harnad, S. (2025). Language writ large: LLMs, ChatGPT, meaning, and understanding. *Frontiers in Artificial Intelligence, Volume 7 - 2024*.

Harris, M. (2023). *Palo Alto: a history of California, capitalism, and the world*. Hachette UK.

Haspel, G., Boyden, E. S., Brown, J., Church, G., Cohen, N., Fang-Yen, C., Flavell, S., Goodman, M. B., Hart, A. C., Hobert, O., et al. (2023). To reverse engineer an entire nervous system. *arXiv preprint arXiv:2308.06578*.

Herbert, F. (1965). *Dune*. Chilton Book Company.

Hurley, M. M., Dennett, D. C., & Adams Jr, R. B. (2013). *Inside jokes: Using humor to reverse-engineer the mind*. MIT press.

Impagliazzo, J., & Proydakov, E. (2011). *Perspectives on Soviet and Russian Computing*. Springer.

Jackson, R. L., Rogers, T. T., & Lambon Ralph, M. A. (2021). Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature human behaviour*, *5*(6), 774–786.

Jankowska, M. A. (1993). Computer technology in eastern European countries and the former Soviet Union: An interpretative bibliography. *Reference Services Review*, *21*(2), 59–76.

Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? (J. Diedrichsen, Ed.). *PLOS Computational Biology*, *13*(1), e1005268.

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, *16*(4), 767–778.

Kerr, R., & Wood, N. (2004). *Science and Civilisation in China Volume 5: Chemistry and Chemical Technology, Part 12, Ceramic Technology*. Cambridge University Press.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in neural information processing systems*, *32*.

Kwisthout, J., Wareham, T., & Van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cogn. Sci.*, *35*(5), 779–784.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lasonen-Aarnio, M., & Littlejohn, C. (Eds.). (2024). *The Routledge Handbook of the Philosophy of Evidence*. Routledge.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of science*, *48*(1), 19–49.

Leahey, T. (2005). Psychology as engineering. In C. E. Erneling & D. M. Johnson (Eds.), *The mind as a scientific object: Between brain and culture*. Oxford University Press.

Lee, N. Y. L., & Johnson-Laird, P. N. (2013). A theory of reverse engineering and its application to Boolean systems. *Journal of Cognitive Psychology*, *25*(4), 365–389.

Legg, T., Hatchard, J., & Gilmore, A. B. (2021). The science for profit model—how and why corporations influence science and the use of science in policy and practice (S. A. Glantz, Ed.). *PLOS ONE*, *16*(6), e0253272.

Leibniz, G. W. (1714). *The monadology*. Springer.

Leonard, A. (2006, January). A twisted tale of Chinese porcelain Reverse engineering, industrial espionage: Been there, done that, got the T-shirt in the 17th century. https://www.salon.com/2006/01/25/porcelain/

Levin, V. K. (2016a). An Essay on forming the Unified System of Electronic Computers (Part I). *translated by Alexander Nitussov. Online accessed on the 11th of November*.

Levin, V. K. (2016b). An Essay on forming the Unified System of Electronic Computers (Part II). *translated by Alexander Nitussov. Online accessed on the 11th of November*.

Levshina, N. (2021). Communicative efficiency and differential case marking: A reverse-engineering approach. *Linguistics Vanguard*, *7*(s3).

Litch, M. (1997). Computation, connectionism and modelling the mind. *Philosophical Psychology*, *10*(3), 357–364.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Malinovsky, B., Fitzpatrick, A., & Aronie, E. (2010). *Pioneers of Soviet computing*.

Marom, S. (2009). On the precarious path of reverse neuroengineering. *Frontiers in Computational Neuroscience*, *3*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

McCarthy, A. M., & Keil, F. C. (2023). A right way to explain? function, mechanism, and the order of explanations. *Cognition*, *238*, 105494.

McCulloch, W. S. (1954). Through the den of the metaphysician. *The British Journal for the Philosophy of Science*, *5*(17), 18–31.

Miłkowski, M. (2013). Reverse-engineering in cognitive-science.

Miłkowski, M. (2016). Computation and multiple realizability. In *Fundamental issues of artificial intelligence* (pp. 29–41). Springer.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of science*, *56*(2), 288–302.

Özgündoğdu, A. F. Ç. (2005). Bone china from Turkey. *Ceramics Technical*, (20), 29–32.

Parberry, I. (2013). Knowledge, understanding, and computational complexity. In *Optimality in biological and artificial networks?* (pp. 125–144). Routledge.

Paul, L., Ullman, T. D., De Freitas, J., & Tenenbaum, J. B. (2023). Reverse-engineering the self.

Petrick, E. R. (2020). Building the black box: Cyberneticians and complex systems. *Science, Technology, & Human Values*, *45*(4), 575–595.

Petrov, V. (2023). *Balkan Cyberia: Cold War Computing, Bulgarian Modernization, and the Information Age Behind the Iron Curtain*. MIT Press.

Pietraszewski, D., & Wertz, A. E. (2011). Reverse engineering the structure of cognitive mechanisms. *Behavioral and Brain Sciences*, *34*(4), 209–209.

Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford University Press.

Polger, T. W., & Shapiro, L. A. (2023). The puzzling resilience of multiple realization. *Minds and Machines*, *33*(2), 321–345.

Poth, N. (2022). Schema-centred unity and process-centred pluralism of the predictive mind. *Minds and Machines*, *32*(3), 433–459.

Quine, W. V. O. (1953). *From a logical point of view*. Harvard University Press.

Rampling, J. M. (2020). *The Experimental Fire: Inventing English Alchemy, 1300–1700*. University of Chicago Press.

Rekoff, M. G. (1985). On reverse engineering. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(2), 244–252.

Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical society*, *74*(2), 358–366.

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science?

Ross, L. N. (2020). Multiple realizability from a causal perspective. *Philosophy of Science*, *87*(4), 640–662.

Ross, L. N., & Bassett, D. S. (2024). Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, *25*(2), 81–90.

Rozemond, M. (1998). *Descartes's dualism*. Harvard University Press.

Rozemond, M. (2014). Mills can't think: Leibniz's approach to the mind-body problem. *Res Philosophica*, *91*(1), 1–28.
Rueckl, J. G. (2012). The limitations of the reverse-engineering approach to cognitive modeling. *Behavioral and Brain Sciences*, *35*(5), 305–305.
Russell, B. (1918). *The philosophy of logical atomism*. Routledge.
Ryle, G. (1931). Systematically misleading expressions. *Proceedings of the Aristotelian society*, *32*, 139–170.
Ryle, G. (2009/1949). *The concept of mind*. Routeledge.
Samuelson, P., & Scotchmer, S. (2001). The law and economics of reverse engineering. *Yale LJ*, *111*, 1575.
Sayward, C. (1983). Minds, substances, and capacities. *Philosophy and Phenomenological Research*, *44*(2), 213–225.
Schierwagen, A. (2012). On reverse engineering in the cognitive and brain sciences. *Natural Computing*, *11*, 141–150.
Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45).
Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.
Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423.
Sejnowski, T., Koch, C., & Churchland, P. (1988). Computational neuroscience. *Science*, *241*(4871), 1299–1306.
Serrano, E., Mercelis, J., & Lykknes, A. (2022). "I am not a Lady, I am a Scientist." Chemistry, Women, and Gender in the Enlightenment and the Era of Professional Science. *Ambix*, *69*(3), 203–220.
Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10).
Simon, H. A. (1995). Artificial intelligence: An empirical science. *Artif. Intell.*, *77*, 95–127. https://api.semanticscholar.org/CorpusID:29911339
Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., Gronau, Q. F., Kalish, M. L., McMullin, S. D., Navarro, D. J., & Shiffrin, R. M. (2022). Statistics in the service of science: Don't let the tail wag the dog. *Computational Brain & Behavior*, *6*(1), 64–83.
Stebbing, S. (2022). *Thinking to some purpose*. Taylor & Francis.
Stengers, I. (2018). *Another science is possible: A manifesto for slow science* (1st ed.). Polity.
Stinson, C. (2018). Explanation and connectionist models. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind* (pp. 120–133). Routledge.
Sullivan, E. (2022). Understanding from machine learning models. *British Journal for the Philosophy of Science*, *73*(1), 109–133.
Szabó, M. (2020). From the West to the East and back again: Hungary and the Ryad. *History of Computing in the Russia, former Soviet Union and Council for Mutual Economic Assistance countries*, 27.
Tenenbaum, J. B. (2021). Reverse-engineering core common sense with the tools of probabilistic programs, game-style simulation engines, and inductive program synthesis. *Proceedings of the Genetic and Evolutionary Computation Conference*.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.
Tennor, M. K. (2015). Reverse engineering cognition. *MITRE CORP MCLEAN VA, Tech. Rep.*
Ter-Ghazaryan, A. (2014). Computers in the USSR: A story of missed opportunities. *Russian beyond*. https://www.rbth.com/science_and_tech/2014/09/24/computers_in_the_ussr_a_story_of_missed_opportunities_40073.html
Vaesen, K., & van Amerongen, M. (2008). Optimality vs. Intent: Limitations of Dennett's Artifact Hermeneutics. *Philosophical Psychology*, *21*(6), 779–797.
van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.
van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697.
van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, *7*, 616–636.
van Rooij, I., & Wareham, T. (2008). Parameterized complexity in cognitive modeling: Foundations, applications and opportunities. *The Computer Journal*, *51*(3), 385–404.
van Rooij, I., & Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. *Journal of Mathematical Psychology*, *56*(4), 232–247.
van der Gun, L., & Guest, O. (2024). Artificial intelligence: Panacea or non-intentional dehumanisation? *Journal of Human-Technology Relations*, *2*.
Vashisth, A., & Kumar, A. (2013). Corporate espionage. *Business Information Review*, *30*(2), 83–90.
Watson, T. J., & Petre, P. (2000). *Father, Son & Co.: My life at IBM and beyond*. Bantam.
Wellman, D. A. (1989). *Chip in the Curtain: Computer Technology in the Soviet Union*. United States Government Printing Office.
Yamins, D. L. (2019). Dan Yamins. https://stanford.edu/~yamins/
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.
Yasuhara, Y. (1991). The myth of free trade: the origins of COCOM 1945–1950. *The Japanese Journal of American Studies*, *4*, 132.
Yoo, S. J. B., Srouji, L. E., Datta, S., Yu, S., Incorvia, J. A., Salleo, A., Sorger, V., Hu, J., Kimerling, L. C., Bouchard, K., Geng, J., Chaudhuri, R., Ranganath, C., & O'Reilly, R. C. (2024). Towards Reverse-Engineering the Brain: Brain-Derived Neuromorphic Computing Approach with Photonic, Electronic, and Ionic Dynamicity in 3D integrated circuits. *ArXiv*, *abs/2403.19724*.
Zednik, C. (2014). Are systems neuroscience explanations mechanistic? https://philsci-archive.pitt.edu/10859/
Zednik, C. (2018). Computational cognitive neuroscience. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind* (pp. 357–396). Routledge.
Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*(12), 3951–3985.

Пржиялковский, В. (1997). Исторический обзор семейства ЕС ЭВМ. https://computer-museum.ru/histussr/es_hist.htm
Савватеев, И. (2023). ЕС ЭВМ. Введение. https://habr.com/ru/articles/732522/